



VALUE-ADDED ASSESSMENT ANNOTATED BIBLIOGRAPHY

Prepared by: Leigh McGuigan
The Ohio State University
October 1, 2003

An updated version will be available October 2006.

This bibliography includes scholarly work¹, from 1993 to the present, which discusses value-added statistical models in the context of K-12 education in the United States². Research studies that rely on value-added models and discussions of value-added assessment in a policy context are included. Works by statisticians concerning the technical features of statistical models that are used in value-added assessment are not included.

Allington, R. (n.d.) The unfairness inherent in value-added assessment of teacher effectiveness. Unpublished manuscript, University of Florida.

Value-added systems do not provide a fair basis for estimating student academic growth, especially in the context of teacher evaluations and pay schemes, because of wide variability in summer reading loss based on socio-economic status, and because of parent funded tutoring and books. Neither of these is adequately controlled for in current value-added models.

¹ Given the use of the Internet to publish academic studies and the increasing conduct of research by private foundations and organizations that do not publish in academic journals, there is no clear definition of what counts as scholarly research. The bibliography includes publications found in academic databases, chapters from academic books and research studies that were apparently performed with scholarly rigor by persons who appeared qualified to undertake such research. All should be evaluated on their own merits.

² There is also significant research in the United Kingdom, where value-added measures are in widespread use.

Baker, A. P., Xu, D. (1995). The measure of education: a review of the Tennessee value-added assessment system. Nashville, TN: Office of Education Accountability, Comptroller of the Treasury.

www.comptroller.state.tn.us/orea/reports

The Office of Education Accountability in Tennessee commissioned this study of the Tennessee Value-Added Assessment System (TVAAS), in order to evaluate the system and identify implementation issues. The authors raised a number of issues, including: unexplained variability in national norm gains across grade levels; large changes in school level scores from year to year; lack of explicit control for confounding factors; and variation in results at one school that seemed to indicate flaws in the system. The authors were unable to come to agreement with Dr. William Sanders, developer of the system, and Sanders' responses are included.

Ballou, D. (2002) Sizing up test scores. *Education Next* 2(2). www.educationnext.org
Practical problems associated with value-added analysis render it potentially unfair for use for accountability purposes. Three significant problems associated with value-added assessment are: (1) testing does not measure gains accurately; (2) factors other than teacher or school quality are related to gain rate; and (3) it is difficult to compare the gains of students at differing ability levels. Thus, there are too many uncertainties to use value-added models for personnel decisions.

Bembry, K.L., Jordan, H.R., Gomez, E., Anderson, M.C., Mendro, R.L. (1998). Policy implications of long-term teacher effects on student achievement. Paper presented at the 1998 annual meeting of the American Educational research Association, San Diego, CA.

Results of the Tennessee Value-Added Assessment System and studies in Dallas are consistent and show highly similar distributions of teacher effectiveness measured by longitudinal analysis of student achievement data. The effects of ineffective teachers are persistent and cannot easily be erased by subsequent assignment to effective teachers. Preliminary findings also suggest that teacher effectiveness measures show little variance over years. The major effect of school leaders in changing the effectiveness of schools comes from changing teaching staff. Districts must find ways to ameliorate the impacts of less-effective teachers and to eliminate the systemic bias in assigning lower performing students to less effective teachers.

Betts, J.R., Zau, A.C. & Rice, L.A. (2003). Determinants of student achievement: new evidence from San Diego. Public Policy Institute of California.

www.ppic.org/main/publication

San Diego schools began a reform program in 2001. This study examined student, teacher and classroom data from 1997 to 2000 to provide statistical estimates of which school and classroom factors have the most influence on achievement growth, and to establish baselines from which reform efforts can be measured. The authors used Stanford 9 tests to measure achievement and estimated regression models to account for achievement gains. Variables with

an important impact on gains included days absent from school and number of peers in the same grade with high scores on the previous test. Class size was important in reading in the elementary grades, but there was no evidence that class size affected achievement gains in higher grades. On the whole, there was no significant difference between fully credentialed and experienced teachers and other teachers, although teacher qualifications became more important in the upper grades.

Bock, R.D., Wolfe, R. & Fisher, T. (1996) A review and analysis of the Tennessee Value-Added Assessment System. Nashville, TN: Comptroller of the Treasury. www.comptroller.state.tn.us/orea/reports/tvaasp1.pdf.

A review of TVAAS by outside reviewers found the system to be fundamentally sound and consistent with other hierarchical models in widespread use. The completeness of the data was found adequate, although improvements could be made. The authors found that teacher effects were stable enough to identify effective and ineffective teachers, but suggested changes that would improve year-to-year instability in school estimates and the accuracy of teacher scores. They question the wisdom of using gain standards based on mean gains reported by McGraw-Hill.

Bryk, A. S., Deabster, P.E., Easton, J.Q., Luppescu, S., & Thum, Y.M. (1994). Measuring achievement gains in the Chicago Public Schools. *Education and Urban Society*, 26(3), 306-319.

Recognizing the limitations of using average test scores to measure changes brought about by school reform efforts, the authors explore a way to assess the value-added by schools in Chicago over and above the contributions of students, families and communities. They begin by rescaling scores from differing versions of the Iowa Test of Basic Skills using Rasch measures, and equated tests both vertically among grades and horizontally across grades. Comparable scores showed some declines in student scores over time. The data showed a counterintuitive “reverse cohort effect” in which average test scores declined in higher grades in years prior to declines in lower grades. The authors plan to model residual school effects using hierarchical linear model analysis. They will compare the baseline pre-reform student gains, adjusted for differences in student populations, against post-reform student gains.

Bryk, A.S., Thum, Y.M., Easton, J.A., & Luppescu, S. (1998). Academic productivity of Chicago Public elementary schools. Chicago: Consortium on Chicago School Research.

In the context of Chicago school reform, the productivity of individual schools cannot be measured fairly using system-wide reports of average test scores. The authors develop a school productivity profile that estimates the value a school adds to learning in a year. The initial study summarizes trends in reading and mathematics achievement, using a new scoring metric that allows comparison of different test content over time. Subsequent studies will examine schools that appear to be especially effective.

California Department of Education, Office of Policy and Evaluation, Special Studies and Evaluation (1998).

The California Department of Education summarizes the value-added accountability models in use in Tennessee and in Dallas, and discusses their potential for California schools. The Tennessee system (TVAAS), was implemented in 1992 with support by the Tennessee Business Roundtable. Three hallmarks of the statistical model are that it controls for confounding variables by allowing each student to serve as his or her own control, it accommodates but does not over-react to missing data, and it protects against misclassification, particularly when there are few scores, by assuming that all teacher or school effects are an average of their school system until the weight of the data pulls their specific estimates away from the mean. Two reviews of the model commissioned by the Tennessee Comptroller have largely confirmed the soundness of the methodology. Despite apparent success by large Tennessee districts in using the TVAAS scores as part of teacher evaluations, policy reviewers continue to oppose using scores for evaluation purposes. The Dallas system, also in use since 1992, uses a two-stage model. Stage one regression controls for confounding influences, and stage two hierarchical linear modeling controls for school level influences. Reviewers suggest that this system may not accurately account for confounding factors, and does not account well for missing data or regression to the mean. A value-added system could be implemented in California.

Calton, J.E. (1995). A study of the understanding and attitudes toward selected parts of the Education Improvement Act of 1992 among Tennessee superintendents and directors of schools. (Doctoral dissertation, Tennessee State University).

The Education Improvement Act (EIA) in Tennessee mandated sweeping educational reforms, including TVAAS. The author included TVAAS in this study of 110 superintendents' attitudes toward the EIA. Only 9.1% of respondents agreed with the statement that "the TVAAS provides a fair, objective means of evaluating the performance of systems, schools and individual teachers," and 29.1% strongly disagreed. Understanding of TVAAS was rated by the respondents as follows: 10% said they had "complete" understanding, 42.7% "good", 33.6% "moderate", 12.7% "little", and .9% "none".

Ceperley, P.E. & Reel, K. (1997) The impetus for the Tennessee value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 133-136). Thousand Oaks, CA: Corwin Press.

TVAAS was at the heart of Tennessee's 1992 education reform legislation, prompted by a funding lawsuit. The Business Roundtable demanded accountability provisions for schools in exchange for business support of tax increases for school funding. The reform plan prescribed penalties for schools and districts that repeatedly failed to make gains equal to national norms. Legislation in 1995 provided that teacher scores were to be included as one of

five items in evaluations. Tennessee Education Association leaders added amendments to weaken the impact of value-added assessments, including the stipulation that teacher scores remain confidential.

Crane, J. (2002) The promise of value-added testing. Progressive Policy Institute. www.ppionline.org

No Child Left Behind brings a new emphasis on accountability. Value-added assessment would provide a better picture than annual yearly progress for measuring progress, because we want to measure gains. Value-added can measure teacher quality and can be used to evaluate reform programs.

Cunningham, L. (1997). In the beginning. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 75-80). Thousand Oaks, CA: Corwin Press.

The Dallas Independent School District built statistical and evaluative capacity throughout the 1960's and 1970's under a remarkably capable superintendent. William Webster, head of research and evaluation, and his staff are leaders in evaluation. Dallas is a leader in educational accountability systems. The value-added assessment system used in Dallas was built over a number of years and consistently refined to fit into (and at times guide) the state accountability scheme.

Darlington, R.B. (1997). The Tennessee value-added assessment system: a challenge to familiar assessment methods. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. pp. 163-168). Thousand Oaks, CA: Corwin Press.

In this evaluation of TVAAS, the author asks two questions: does the system work well enough to be worthwhile; and how does it compare to regression methods. He concludes that the system does work well enough, and that it is superior to regression models in most real world contexts, where significant data is missing. Regression models are well understood by educators, and thus would be preferable in an ideal school system with complete data, but no such system exists. Mixed models are not well known in education circles, being more common in agricultural statistics. Nonetheless, the mixed model seems superior because of its handling of missing data. TVAAS can be an important tool in evaluating teachers, although subjective judgment must also be applied.

Doran, H. & Drury, D. (2002). Evaluating success: Kipp educational program evaluation. New American Schools technical report.

A New American Schools report evaluates three charter schools, measured by academic achievement gains. Gain rates at the schools showed that among all subgroups, students progressed fastest after enrolling at the schools.

Drury, D. & Doran, H. (2003). The value of value-added analysis. *National School Boards' Association Policy Research Briefs* 3(1). www.nsba.org

Value-added analysis is an excellent complement to traditional means of measuring school effectiveness. There are three prerequisites for implementing a value-added system: annual testing; test scores reported on a common scale;

and the ability to uniquely identify students. Concerns expressed include lack of transparency and the possibility of random error; however, we cannot sacrifice rigor for transparency, and random error can be minimized through averaging and multiple measures. Value-added data, combined with other measures, can inform policy and program evaluation, motivate instructional change, serve as a tool for evaluating teacher effectiveness, and guide individualized professional development. Value-added models can be reconciled with the average yearly progress requirements of No Child Left Behind.

Goycochea, B.B. (1998). Rich school, poor school. *Educational Leadership* 55(4), 30-33.

One-time average scores on a writing test did not represent the full story at a California middle school. After implementing a new writing program to improve student skills, the school still found itself graded as failing based on student averages. Inspired by Webster and Mendro's value-added approach in Dallas, the school calculated student level gains compared to those at the same district's flagship school. Results showed that, controlling for beginning language ability, the schools produced comparable gains. This finding increased teacher and administrator morale and altered attitudes about the school's potential for success.

Greene, J.P. (2002) The business model. *Education Next* 2(2). www.educationnext.org
Flaws in standardized tests generally, and in value-added models in particular do not lead to the conclusion that they should not be used. The benefits of testing, and in value-added models, outweigh their drawbacks. Schools must measure productivity so that it can be rewarded. Between 1961 and 2000, after adjusting for inflation, education spending in the United States tripled while performance measured by the NAEP remained flat. At the same time as educational productivity was stagnating, however, private businesses achieved steady productivity growth because competition forced the adoption of productivity measurement systems. These are not flawless, either, but they have proven to work. As long as errors in productivity measurement and reward systems are mostly random and do not create perverse incentives, even flawed systems are effective. Value-added data could revolutionize systems for hiring, promoting and compensating teachers. The balance in education must shift toward the mechanical application of measurement systems, because principals and other administrators have few incentives (unlike their private sector counterparts) to make appropriate subjective judgments.

Heck, R. (2000). Examining the impact of school quality on school outcomes and improvement: a value-added approach. *Educational Administration Quarterly* 36(4), 513-52.

For accountability systems to be fair, state report cards should take into account differences in student populations. This approach to statewide school comparison in Hawaii looks at the value-added by the school, adjusted for student population differences. Some use a "school effect residual" that represents the difference between observed scores and scores that can be

predicted from variables that describe student populations. The author advocates multi-level modeling, which allows researchers to assign explanations to various levels of a data hierarchy such as the school, student, teacher or district. The model includes student variables such as prior achievement, gender, ethnicity and language status, plus school context variables such as leadership, academic emphasis and school climate. This model can then explore which characteristics of schools explain improvements in student gains.

Holloway, J.H. (2000). A value-added view of pupil performance. *Educational Leadership* 57(5), 84-86.

The author reviews how several practitioners of value-added analysis account for all the non-school variables that influence achievement, citing Tennessee findings that the largest factor in student gain is the teacher. He cites Bryk, Thum, et al. (1998) in pointing out that when tests change over time, there is no absolute scale of growth. Teachers, for example, sometimes see lower test gains reported even among student populations that they believe have higher knowledge. This possibility should be accounted for. In addition, student mobility affects averages. Data must be analyzed at the individual student level. In a California district, a value-added analysis was used by a middle school with a large proportion of poorer students to show that, after implementation of a writing program, their students' gains were comparable to the district's wealthy flagship school.

Herdman, P., Smith, N. & Skinner, C. (2002). Charter Schools and the New Federal Accountability Provisions. New American Schools, Education Performance Network, 675 N. Washington St., Suite 220, Alexandria, VA 22314.

Izumi, L.T. & Evers, W.M. (2002). In Evers, W.M. & Wahlberg, H.J. (Eds.) *School Accountability* (pp. 105-153). Stanford, CA: Hoover Institution Press.

The Texas and Florida accountability systems, which both include a value-added component, are described in an overall chapter about accountability. In the Texas system, the Texas Learning Index measures growth by comparing student achievement gains in each school to those of a group of similar schools. Schools are ranked in quartiles. Bonuses ranging from \$500 to \$1500 are awarded to teachers in schools ranked in the top quartile. Low-performing schools are subject to sanctions. In Florida, the Sanders value-added model is used to provide a component of school scores on a state report card. In 1997 (amended in 1999) the Florida legislature required school boards to base a portion of each employee's compensation on student performance. The weight to be given to student performance at the classroom level is left unspecified.

Kupermintz, H, Shepard, L. & Linn, R. (2001). Teacher effects as a measure of teacher effectiveness: construct validity considerations in TVAAS (Tennessee Value-Added Assessment System). In D. Koretz (Chair), *New work on the evaluation of high-stakes testing programs*. Symposium at the National Council on Measurement in Education (NCME) Annual Meeting, Seattle, WA. April, 2001.

The Tennessee value-added model claims that “blocking” by using each student as a unit partials out all effects of socio-economic status. This raises two concerns: first, that prior achievement does not completely control for all socio-economic and other factors; and second, that there is a confounding of student achievement and teacher effectiveness, because the “treatment”, that is, teacher effectiveness, is not statistically independent from the student prior achievement. The author demonstrates the confounding effect using a simulation. Because the model assumes teachers perform at the mean when student data is insufficient, it is nearly impossible for teachers with few students to differ measurably from the mean. These concerns result in the unequal treatment of teachers. The authors view the Tennessee model as “narrow” and “mechanistic”, and dispute the conclusion that student performance gains can be equated with effective teaching.

Kuppermintz, H. (2002). Value-added assessment of teachers: the empirical evidence. In Molnar, A. (Ed.) *School reform proposals: the research evidence*. Greenwich, CT: Information Age Publishing.

TVAAS raises a number of questions, including: do teacher effect scores capture teachers’ contributions to student learning; do scores provide equal standards of excellence for all teachers; do TVAAS scores reflect desirable instructional practices; and are student test scores adequate measures of the desired outcome of teaching. Both the dynamic nature of learning and the many influences on the process of learning impede research. Flaws in the Sanders model include incomplete experimental control in not accounting for confounding factors, and the confounding of student achievement and teacher effectiveness. There are also significant issues of accuracy of the model. In sum, the model is narrow and mechanistic, and ignores process.

Ladd, H., & Walsh, R. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review* 21(1) pp. 1-17.

Value-added systems must distinguish the effectiveness with which schools operate, controlling for differing resources. North and South Carolina school effectiveness measures are based on overall gains in test scores, and do not accurately account for how efficiently a school is operated because they include effects outside the control of school personnel. While these measures provide useful information, because they do not control for school resources, including the student mix, that are outside the control of schools, they should be used cautiously for accountability purposes. Results show that schools serving higher performing students are more likely to be deemed effective than schools serving lower performing students. The authors determine that two-fifths of this correlation is due to measurement error. However, even correcting for measurement error, effectiveness remains correlated with socio-economic status and average test scores. The likely result is that the best teachers will be drawn to the better performing schools.

Ladd, H.F. (1999). The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes. *Economics of Education Review* 18, pp.1-16.

Gains in student performance in Dallas, where value-added assessment is in use, are compared to those in other cities. Hispanic and white seventh graders in Dallas appear to have significantly greater achievement, but the same result is not seen in black students.

Lee, K. & Weimer, D. (2002). Building value-added assessment into Michigan's accountability system: lessons from other states. The Education Policy Center, Michigan State University. www.epc.msu.edu

The authors describe value-added evaluation schemes in use in Dallas, in Texas as a whole, in Tennessee and in North Carolina. The Dallas model employs regression and hierarchical linear modeling techniques, using longitudinal student achievement data. The Tennessee model, developed by Dr. William Sanders, is based on individual student test scores, and excludes variables such as socio-economic status. Each student's gains are compared to his or her own performance over a three-year period. The Texas model calculates annual gains by subject, by student. These gains are then averaged to find school gains. Schools are then compared to similar schools. The North Carolina model measure schools based on an expected growth rate for the year. The expected growth of a cohort is based on an average annual growth rate for the state's students, controlling for previous achievement and student proficiency. Possibilities for implementing a value-added model in Michigan are explored.

Lockwood, J.R., Louis, T. & McCaffrey, D. (2002). Uncertainty in rank estimation: implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics* 27(3), 255-70.

The performance of rank (percentile) estimators is investigated in a basic, two-stage hierarchical model that captures the essential features of value-added systems currently in use. The authors conclude that percentile estimators do not necessarily perform well enough to be desirable for use as a basis for evaluations. Substantial information, which is not always present, is needed for acceptable performance of systems based on percentile rankings.

Logan, M.C. (1995). A study of Davidson County's elementary and middle schools: teachers' perceptions of the Tennessee Value-Added Assessment System. (Doctoral dissertation, Tennessee State University.)

In 1994, the author sent a questionnaire to teachers at 25 randomly selected middle and elementary schools, receiving 252 responses. Fifty percent of respondents agreed with the statement that TVAAS was "a waste of money." Only approximately 21% agreed with the statement that they "agreed with or supported" TVAAS. A significant majority disagreed with the statement that TVAAS had been thoroughly explained, and significant majorities agreed with statements that more information was needed about TVAAS and that the emphasis on teacher effects was too great. 122 out of 167 strongly disagreed that TVAAS gains were true indicators of teacher effects, even though 181 out of

242 agreed that test gains accurately reflect teacher effectiveness. 135 out of 246 disagreed that "TVAAS is a fair system," and 117 out of 241 disagreed with the statement "I am supportive of TVAAS."

McAdams, D. (2002). Enemy of the good. *Education Next* 2(2)
www.educationnext.org

Even though value-added analysis is flawed, the author, a former Houston school board member, is committed to it. It is better than the other measures. However, it must be applied with flexibility and supervisors must have the freedom to use their judgment. High-performing organizations measure almost everything, because measurement changes behavior. Constant measurement focuses attention. While it is obvious that testing can be misused, there is little evidence of value-added being misused, and much evidence that it focuses attention on learning.

Mendro, R.L., Jordan, H.R., Gomez, E., Anderson, M.C., and Bembry, K. (1998). An application of multiple linear regression in determining longitudinal teacher effectiveness. Paper presented at the 1998 annual meeting of the American Educational Research Association, San Diego, CA.

Building on work by Raudenbush and Bryk and by Sanders, Dallas District research staff conducted an analysis of teacher effectiveness in Dallas using multiple linear regression. They discuss the design and implementation of a hierarchical linear model to analyze longitudinal teacher effectiveness data over three years.

Meyer, R. (2002). Value-added indicators: do they make an important difference? Evidence from the Milwaukee public schools. Wisconsin Center for Education Research. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 2, 2002.

The Milwaukee Public Schools initiated annual assessment of students in 2001 and is now using value-added analysis to track school performance, program efficacy, and other policies. Math achievement data from 7th and 8th grades is used to illustrate the approach. The purpose of using value-added analysis is to isolate the contribution of schools from other factors, such as prior achievement and student, family and neighborhood characteristics. The value-added math productivity of Milwaukee middle schools varied widely, with some generating almost two times the growth of the average school. Most schools the served students with lower prior achievement produced better than average gains.

Meyer, R.H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review* 16(3), 283-301.

Meyer, R.H. (1996). Value-added indicators of student performance. In Hanushek, E.A. & Jorgenson, D.W. (Eds.). *Improving America's Schools: The Role of Incentives* (pp. 197-224). Washington: National Academy Press.

The author's premise is that level indicators such as average test scores, while valid for some purposes, are not valid measures of school performance. He

describes various types of statistical models that can be used to determine gains in student achievement, discusses the contexts in which particular models can be effective, and examines the factors that affect reliability. He compares value-added assessments to other, more common, indicators of student achievement, and suggests that inappropriate indicators of achievement drive flawed policy decisions. He recommends that students be tested at every grade level and that other data be collected to enable value-added assessments of student gains.

O'Brien, D. & Ware, A. (2002). Implementing research-based reading programs in Fort Worth Independent School District. *Journal of Education for Students Placed at Risk* 7(2), 167-95.

The Fort Worth Independent School district implemented a direct instruction reading program in kindergarten, first and second grades in 61 schools. Two different programs were implemented. This article reports on the history, implementation and evaluation of the programs. Outcome evaluation used statistical analysis to look at the effect of the program on district performance, to compare schools, programs and classrooms, and to identify sources of variation in student gain. Student gains were evaluated using a value-added regression methodology. Simple comparisons of the two programs with a control group would not be adequate because of large differences in the demographics of school populations. The student level regression analysis used a spring reading test score as the independent variable. Dependent variables included the fall reading test score, race, free or reduced price lunch status, gender, special education status and English proficiency. In 50 of 60 program years, holding control variables constant, score gains for the direct instruction programs were larger than the control school gains.

Patteson, F.E. (2002). Educational values and accountability in Tennessee: ethical dilemmas and moral imperatives. PhD. dissertation, University of Tennessee.

A case study involving 60 participants, including teachers, principals, state officials, politicians and students, examined perceptions of the Tennessee accountability system. Historical criticisms of the accountability system are summarized, and the author recounts two instances where TVAAS was a partial basis for disciplinary action, including the demotion of a school director. Among the 60 subjects of the study, the majority of subjects did not think that the accountability system's sole reliance on test scores was appropriate. TVAAS was criticized as a basis for teacher accountability because some teachers did not get scores.

Peevely, G. & Ray, J. (2001). Does equalization litigation effect a narrowing of the gap of value-added achievement outcomes among school districts? *Journal of Education Finance* 26(4), 463-76.

Value-added student achievement levels of the mostly small, rural schools that participated in the Tennessee school finance case were compared to districts that had not been litigants. The litigants asserted that inequitable funding resulted in lower achievement levels, and implied that higher expectations required equalization of resources. When "full funding" as mandated by the state

was achieved, litigants budgets increased, but were still less than other districts. During the phase in of full funding, the gap between litigants and non-litigants did not significantly close, and the direction of change was mixed.

Phillips, B. (2002) The use of value-added gain scores to assess the impact of school funding. *Education Leadership Review* 3(2), 17-21.

Public Education Foundation (n.d.). Teacher Quality Initiative (unpublished)
www.perfchattanooga.org

The authors report the preliminary results of a study of highly effective teachers in 92 elementary and middle schools in Hamilton County, Tennessee. The authors are seeking to determine the qualities of teachers who have been rated effective by the TVAAS system, or, if they do not have a rating, by nomination of their principals. Each teacher participated in interviews, surveys and personality inventories and allowed classroom visits and personnel file reviews. The effective teachers are diverse, but they defined an effective teacher similarly, using terms such as flexible, good manager, caring, and having high expectations. Observations showed that the teachers displayed large amounts of student work, covered the room as they taught rather than standing still, used multiple small group activities, put students at ease, were organized, and had clear, high expectations. Seventy percent of the sample had a family tradition of educators.

Rivers, J. & Sanders, W. (2002). Teacher quality and equity in educational opportunity: findings and policy implications. In Izumi, L.T. & Evers, W.M. (Eds.) *Teacher Quality*. CA: Hoover Press.

The authors define educational equity as “each student making appropriate academic growth each year.” Academic growth that is within control of educators depends on the district, the school, but most importantly on the teacher. The teacher effect is far greater than classroom variables such as previous achievement of the class, class size, and socioeconomic status of the class. Six million student records that have been analyzed as part of the TVAAS show that there is extreme variability among teachers that dramatically affects student academic growth. This variability increases with grade level and is most pronounced in mathematics. For example, fifth graders who had highly ineffective teachers in grades three through five scored 50 percentile points below peers of comparable previous achievement levels who had highly effective teachers. Teachers’ effects on student achievement persisted at least four years after the student no longer had the teacher. The effect of teachers is separate from socioeconomic status, ethnicity, and parental education. Regardless of ethnicity, children of similar previous achievement levels respond similarly to a given teacher. In light of these findings, policy makers must continue to measure effects, shrink the variability among teachers, and find ways to minimize the impact of ineffective teachers.

Roderick, M, Bryk, A., & Jacob, B. (2002) The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation & Policy Analysis* 24(4), 333-57.

Chicago has linked school accountability with high stakes consequences for students by requiring mandatory levels of performance on standardized tests to achieve promotion from 3rd, 6th and 8th grades. Using a three-level hierarchical linear model, the authors compare student achievement gains in grades targeted for promotion both before and after implementation of the policy, and conclude that student achievement increased substantially in grades 6 and 8 after implementation of the policy. Results for grade 3 were less conclusive. In reading, the lowest achieving students showed the greatest gains, while in math the opposite was true.

Ross, S.M., Stringfield, S., Sanders, W.L., Wright, S.P. (2003). Inside systemic elementary school reform: teacher effects and teacher mobility. *School Effectiveness & School Improvement* 14(1), 73-110.

Although many studies of school reform efforts focus on at the school level, it is important to focus on teachers. This study examines the effects of Memphis school reform efforts on teachers' effectiveness. In the 1995-96 and 1996-97 school years Memphis implemented Comprehensive School Reform using various reform models. Teacher effectiveness scores were derived from value-added student scores by measuring each teacher's contribution to student gains relative to an average teacher in the district. These scores were then compared for the two reform cohorts and a control group of non-reform schools. While the teachers in the 1995-96 cohort of reform schools significantly outperformed non-reform teachers in producing student gains, this did not hold true with respect to the second reform cohort. In the 1995-96 cohort, highly and moderately experienced teachers became much more effective after reform. This also was not the case in the second cohort. The authors also looked at programmatic differences in student gain outcomes, and at effects of reform on teacher mobility, with mixed findings.

Ross, S.M. (2001). Creating critical mass for restructuring. Appalachia Educational Lab., Charleston, WV. [AWT03075] Office of Educational Research and Improvement, Washington, DC. [EDD00036]

From 1995 to 1999, Memphis city schools implemented a successful systemic reform program. However, in 2001, the program was discontinued. The author documents the success of the program in its early stages, as evidenced by observations and by gains in value-added scores. As the reform program was expanded in its later years, however, schools were added that were more resistant to change and support resources became strained. Schools that joined the reform effort later reported feeling pressure to adopt "favored" reform models. In addition, the superintendent and assistant superintendent who had spearheaded the program announced their departures. The author concludes that for any reform effort to be successful, it must be tailored to school needs, have teacher buy-in, be supportable by district staff, originate in schools rather than in district offices, and be backed by credible, high-quality training.

Ross, S.M., Sanders, W., Wright, S., Stringfield, S., Wang, L., Weiping, A., & Alberg, M. (2001). Two- and three-year achievement results from the Memphis restructuring initiative. *School Effectiveness and School Improvement* 12(3), 323-46.

This study examined the results of three years of school reform initiatives in Memphis. The authors recount the history of recent Memphis reform initiatives and summarize the various reform models chosen by certain Memphis schools. Using TVAAS scores, the study compares the percent of expected gains attained by reform schools versus non-reform schools. After three years, reform schools generally gained relative to matched non-reform schools. Stronger results were shown in the first group of schools to adopt reform models.

Sanders, W. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education* 14(4), 329-39.

In this transcript of a speech given on receiving an award named for Dr. Jay Milman, the author describes TVAAS and addresses common criticisms of value-added assessment. He describes the pattern, especially in inner city schools, in which previously lower scoring students make progress, but earlier higher scoring students fail to do so. The goal should be realistic growth for all students. While measurement alone will not bring change, information has to be available to guide development and growth.

Sanders, W. (1998). Value-added assessment. *School Administrator* 11(3), 24-27.

“Value-added” methodology is defined as “a statistical method of determining the effectiveness of school systems, schools and teachers in sustaining academic growth for student populations.” This methodology allows each student to serve as his or her own control, rather than explicitly accounting for race and socio-economic status. The validity of this approach is confirmed by the fact that demonstrated cumulative school gains are unrelated to racial and socioeconomic factors. The single largest factor affecting academic growth is the classroom teacher, and teacher effects are cumulative and residual. The data show that there are large losses in academic achievement gains when student populations change buildings (such as in middle school transition). In general, students at the highest levels of academic achievement show the smallest gains.

Sanders, W. L. & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: implications for educational evaluation and research. *Journal of Personnel Evaluation in Education* 12(3), 247-56.

Research using the massive longitudinal database in Tennessee shows that teacher effectiveness is a predominant factor in determining student achievement. Analysis of TVAAS data shows a number of patterns, including the negative effects of building change; the cumulative and additive effects of ineffective teaching; that students at the highest achievement levels show less growth than other students; and that black students are disproportionately assigned to ineffective teachers.

Sanders, W. L. & Horn, S.P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education* 8(3), 299-311.

TVAAS is described, including the background for its development and the advantages of using a statistical mixed model approach. The authors defend the practice of using each student as his or her own control rather than trying to explicitly account for all other variables that affect student achievement, noting that it is impossible to account for everything that might have an effect on a student's achievement in a given year.

Sanders, W., Saxton, A., & Horn, S. P (1997). The Tennessee Value-Added Assessment System: a quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

TVAAS is described by its creator as a "repeated measures, multivariate response analysis" that allows the inclusion of all the information for each student regardless of the degree of missing information. It uses each student as his or her own control, because it would be a "hopeless impossibility" to include all data for each child so that all confounding influences could be accounted for. The history of TVAAS implementation in Tennessee and the efforts made to educate teachers and other groups is summarized. Early users of the data included teachers who calculated their own effectiveness based on student scores even before such information was provided by the system. Early findings included significant negative effects of building changes such as from 5th grade to middle school. The model's technical aspects are described, including how it accounts for missing data. The handling of missing data is not unfair to teachers with little data, because it is "nearly impossible" for teachers with little data to have estimates that are measurably different from the mean. The computing requirements of TVAAS are significant.

Sanders, W., Wright, P., Ross, S. & Wang, W. (2000). Value-added achievement results for three cohorts of roots and wings schools in Memphis: 1995-1999 Outcomes. www.successforall.com.

This is an extension of a prior study of Roots & Wings schools in Memphis. Twenty-two Roots & Wings schools were examined, with four, three and two years of implementation history. Among the cohort with the longest history of implementation, achievement gains measured by the TVAAS (using the percent of national norm gain across all grades) significantly exceeded comparison non-reform schools. The results were positive, but less pronounced for cohorts with fewer years of reform experience.

Schacter, J., Schiff, T. Thum, Y.M., Fagnano, C., Bendotti, M., Solmon, L., Firetag, K. and Milken, L. (n.d.). The impact of the teacher advancement program on student achievement, teacher attitudes, and job satisfaction. Milken Family Foundation. www.mff.org/pubs/impact_of_tap.pdf

This study compared gains in student achievement of schools implementing a reform model developed by the Milken foundation to similar schools not implementing the reform model. Student level scores for Stanford 9 Total Reading, Language and Mathematics were used in a value-added model that measured the percent of the gap to a specified target achievement level that was in fact achieved. All reform schools made improvement and gained significantly more than control schools. Those schools that implemented the reforms the most zealously gained the most.

Seltzer, M., Choi, D., & Thum, Y.M. (2002). Examining relationships between where students start and how rapidly they progress: implications for conducting analyses that help illuminate the distribution of achievement within schools. National Center for Research on Evaluation, Standards and Student Testing (CRESST), University of California, Los Angeles.

Although there is an increasing emphasis on monitoring rates of learning change for cohorts of students, it is important, too, to look at the relationship between starting levels of accomplishment and the rate of progress. Analysis shows that comparing rates of change for various demographic groups can be misleading, as size and direction of change, even within the group, may differ markedly depending on the starting point.

Sterbin, A. (2001). Rozelle Elementary School: a longitudinal analysis 1995-2000. University of Memphis. www.mrsh.org/ipr.html

In a case study of one Memphis elementary school, the author used value-added scores, among other factors, to assess the success of Modern Red School House comprehensive school reform model as implemented in the school. Comparing longitudinal value-added scores from the subject school with those of similar schools, all Memphis schools, and all Tennessee schools, he found that value-added scores rose, especially in the early years of reform implementation, far more than in all comparison groups, and remained higher throughout the study than comparable schools and all Memphis schools.

Stone, J. E. (1999). Value-added assessment: an accountability revolution. In Kanstoroom, M., & Finn, C.E., Jr. (eds.). *Better Teachers, Better Schools*. Washington, D.C.: Thomas B. Fordham Foundation.

Value-added assessment systems in use in Tennessee and Dallas are briefly described. While the quality of teaching has in the past been measured by inputs such as teacher licensure, degrees, training, etc., now teacher quality can be directly measured through value-added models. Value-added assessment could be used in teacher licensure, tenure and merit pay decisions, and in measuring curricula and teacher training effectiveness. Value-added assessment has broad appeal to parents and politicians because they are different from educators. While educators view measures of academic achievement as “one outcome among many”, parents and politicians view it as the “indispensable core” of schooling.

Stone, J.E. (2002). The value-added achievement gains of NBPTS-Certified Teachers in Tennessee: a brief report. Thomas B. Fordham Foundation. (Unpublished manuscript). www.education-consumers.com/briefs/stoneNBPTS.shtml

Sixteen of the forty teachers in Tennessee who have been certified by the National Board for Professional Teaching Standards teach in the 3rd to 8th grades and thus have TVAAS scores. An examination of those scores shows that none of the certified teachers achieved the threshold for effectiveness that would qualify them for bonuses under the Chattanooga bonus system, which is an effectiveness score of 115% of the norm. The author concludes that on the basis of TVAAS scores, certified teachers cannot be considered exceptionally effective.

Stronge, J. & Tucker, P. (2000). *Teacher Evaluation and Student Achievement*. Annapolis Junction, MD: NEA Professional Library.

Summers, A. (2002) Expert measures. *Education Next* 2(2). www.educationnext.org
Critics of value-added systems do not want them used for accountability. But it is clear that teachers are the dominant input in schools, but both in terms of spending and in terms of learning. Confidence in value-added models can be developed by replication and by averaging over time periods. Value-added also needs to be coupled with other measures for an effective accountability system. Even if there are flaws, value-added measures are clearly superior to basing teacher compensation on degrees earned or other traditional systems. In sum, the problems with value-added systems are exaggerated, and the alternatives are untenable.

Thum, Y.M. (2003). No Child Left Behind: Methodological challenges & recommendations for measuring adequate yearly progress. CSE Tech Report 590, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

A statistical accountability measure that focuses on value-added performance can be used to measure Annual Yearly Progress (AYP) under NCLB. AYP can be conceived as a comparison of a unit's growth rate to the rate expected if the unit is to reach proficiency by the NCLB deadline. While there are various types of measures called value-added, only a model based on student level gain gives us a true map of change. (Estimating gains scores is preferable to regressing post-test scores on pre-test scores.) Even when the unit of accountability is the school, the unit of analysis should be the student. Reports should show, and accountability regimes should take account of, the precision of the productivity estimate.

Thum, Y.M. (2002). Measuring student and school progress with the California API. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

The California Public Schools Accountability Act required the state board of education to establish a numerical index (API) for measuring school performance and the performance of groups of students within schools. The author concludes that the API misrepresents student and school performance and presents an

alternative based on a Bayesian meta-analysis of results from multilevel models of student test scores. This new approach retains the main features of the API and is feasible given the current API database. The approach is illustrated using data from the Long Beach Unified School District.

Thum, Y.M. (2002). Measuring progress towards a goal: estimating teacher productivity using a multivariate multilevel model for value-added analysis. Milken Family Foundation. www.mff.org

The author presents a general procedure for measuring teacher or school effectiveness based on the relationship between estimated gains and an estimated target gain score. The model accounts for correlated error in multiple measures, simultaneously estimates initial status and gains, and produces a productivity index that includes a representation of the uncertainty in individual estimates. The model is broadly useful for longitudinal analysis and is applicable to a broad range of assessments. It is not limited to use with test scores.

Thum, Y.M. & Bryk, A.S. (1997). Value-added productivity indicators. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Dallas is a leader in school accountability, especially notable for the degree of political support and educator buy in for the system. While the system rules work reasonably well at controlling for student mobility and preventing manipulation, and the adjustment for fairness variables is appropriate, there are significant questions about the validity of the model and it should be used with caution.

Topping, K.J. & Sanders, W.L. (2000). Teacher effectiveness and computer assessment of reading. *School Effectiveness & School Improvement* 11(3), 305-37.

Accelerated Reader (AR) is a computer-assisted assessment, designed to measure reading comprehension that is voluntarily performed by students. It provides feedback on the number and type of books read and on comprehension levels. Using data from Tennessee, the authors merged TVAAS scores with AR data in order to explore changes in teacher effectiveness in reading that might be caused by teacher management of the quality and quantity of reading practice. The authors found a significant relationship between reading volume and comprehension, as measured by AR, and teacher effectiveness, measured by TVAAS scores. In third and fourth grades, TVAAS scores rose with the quality and quantity of books read. This remained true in the fifth and sixth grades, except for a pronounced negative impact when students read very many very easy books. There was a positive relationship between reading comprehension and TVAAS scores across all grades.

Tucker, P. & Stronge, J. (2001). Measure for measure. *The American School Board Journal* 188(9), 34-7.

Teachers are the single most important school influence on learning and teacher effects should be measured. Using value-added assessment scores in teacher evaluation is controversial and is opposed by the National Education Association (NEA). The NEA commissioned the authors to study value-added assessment

systems in Tennessee, Texas, Oregon and Colorado. Although NEA members voted against using value-added measures in evaluating teachers, the authors believe that value-added assessment can be a valuable teacher evaluation tool. Tennessee has the most effective approach, because value-added scores are one of a number of performance indicators considered and because professional development is left to school discretion.

Yu, L. (2001). Measuring value-added school effects on Ohio 6th grade proficiency test results using two level hierarchical linear modeling. (Doctoral dissertation, University of Toledo).

The author provides a detailed history of the development of school effectiveness research, including research in the United Kingdom. Using a two level hierarchical linear model that consisted of the student-level variables eligibility for free or reduced price lunch and prior achievement, and the school-level variables percent of students eligible for free or reduced price lunch, percent of teachers with masters and percent of students suspended, he analyzed the achievement of 1915 6th graders in 44 elementary schools in a Northeast Ohio district, based on scores on the Ohio proficiency test. Schools were shown to account for ten to fourteen percent of the variance in achievement gain.

Vaughan, A. (2002). Standards, accountability, and the determination of school success. *The Educational Forum* 66(3), 206-13.

A review of the history of the modern standards movement from 1989 to the present includes examples of state successes in raising student achievement and narrowing the achievement gap. Policy makers now realize that evaluations based on average test scores simply result in rewarding prosperous schools. Value-added evaluation models such as those in place in North Carolina, Tennessee, Prince Georges County, Maryland, Colonial School District in Pennsylvania, and Dallas, Texas may be the best hope for transforming schools.

Walberg, H.J. & Paik, S.J. (1997). Assessment requires incentives to add value. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

While TVAAS is “driven by technicalities and lacks adequate incentives”, it is the “least bad” of all systems. The designers may have been cowed by teachers unions, which represent one of the reasons that the most productive country in the developed world has the least productive educational system. When private organizations assess performance, they make adjustments to try to arrive at the “value-added” aspects of the performance. However, they don’t wait for the perfect system. The TVAAS lacks “teeth” because there are not “clear and appropriately enacted probation and dismissal policies as well as individually based merit pay.” TVAAS appears to comply for the most part with the standards of the Joint Committee on Standards for Educational Evaluation.

Webster, W.J. (1998). A comprehensive system for the evaluation of schools. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA., April 13-17.

The author describes the three-tier accountability system in place in the Dallas public schools. School effectiveness indices are one component of the system. The inclusion of these indices results in a valid and fair way to compare the effects of the school on student learning, apart from factors out of the school's control.

Webster, W.J., & Mendro, R.L., Orsak, T.H. & Weeransinghe, D. (1998). An application of hierarchical linear modeling to the estimation of school and teacher effect. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA April 13-17.

Criteria for defining and for judging statistical models for estimating school and teacher effects on student learning are discussed and the origin and development of the accountability system developed by the Dallas district's research department are presented.

Webster, W.J., & Mendro, R.L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.

Dallas schools have used a value-added assessment system since 1992, expanded to include teacher effects in 1994. The model uses a combination of multiple regression and hierarchical linear modeling. It is part of an overall accountability system based on district improvement plans, school improvement plans, principal and teacher evaluations and school and teacher effectiveness measures. The model uses regression at the student level to control for the "fairness variables" of ethnicity, gender, language proficiency, socioeconomic status and prior achievement. This regression step, by explicitly taking fairness variables into account, is very important in demonstrating to educators and to the public the fairness of the model. The model uses hierarchical linear modeling at the school level to control for mobility, crowding, percent minority and socioeconomic status. Teacher effectiveness indices provide a valuable tool for professional development. An accountability task force composed of parents, teachers, principals and community and business leaders is the final authority on selecting variables and weightings, the rules of the system, performance awards, and appeals of system decisions. Dallas awarded \$2.4 million to staffs of effective schools (prorated among staff) in each year from 1992-1994, and has raised the amount to \$3 million in 1995. School level awards encourage cooperation and reduce competition among teachers.

Webster, W.J. & Mendro, R.L. (1995). Evaluation for improved school level decision-making and productivity. *Studies in Educational Evaluation* 21, 361-399.

Dallas' three-tier accountability system couples accountability at the school and district levels with school accountability indices. School and district improvement plans support the first two components of the accountability system. The school improvement indices used by Dallas schools separates school effects from non-

school effects in a more precise way than measures used by the Texas Education Agency. The only fair method for holding schools accountable for improvement is to adjust measures of outcomes for factors that impact those outcomes but are out of school control. This can be done through multiple regression and through hierarchical linear modeling. The indices predict student levels of accomplishment and set the desired level of improvement based on these predictions. School-level consequences are imposed based on achievement or non-achievement of predicted levels of student improvement.

Wright, S.P., Horn, S.P. & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11, pp. 57-67.

This study looked at the size of teacher effects on student achievement, controlling for intra-classroom homogeneity, prior achievement level, and class size. Teacher effects were found to be the predominant factor in student achievement gains, with wide variation among teachers. The achievement level of students was also a factor. It was a common pattern for the highest achievers to make the lowest gains. The authors conclude that more could be accomplished by raising teacher effectiveness than by affecting any other single factor in the schools.